Range Mgmt. & Agroforestry 37 (1) : 21-26, 2016
ISSN 0971-2070

Comparison of logistic regression and machine learning techniques in prediction of habitat distribution of plant species

Hossein Piri Sahragard[1] and Mohammad Ali Zare Chahouki[2*]

[1]Range and Watershed Department, University of Zabol, Iran

[2]Department of Rehabilitation of Arid and Mountainous Regions, University of Tehran, Iran

*Corresponding author e-mail: mazare@ut.ac.ir

Received: 17th July, 2015 Accepted: 9th March, 2016

Abstract

The study was carried out to compare performance of Logistic regression (LR) and machine learning techniques to predict habitat distribution of plant species in rangelands of Qom Province, Iran. After determination of homogeneous units, vegetation sampling was carried out using random systematic method. The plot size was determined using minimal area method from 2 to 25 m^2. For soil sampling, at each habitat, eight holes were drilled and samples were taken from 0-30 and 30-80 depths. Soil characteristics consisting gravel percent, texture, saturation moisture, available water, lime, gypsum, organic matter, acidity (pH), electrical conductivity (EC) were measured by standard methods. Using geostatistical and kriging interpolation method with the same spatial resolution soil digital layers were prepared and stored in GIS. Digital elevation map of the region was used for mapping slope, aspect and elevation. After implementation of the models, to evaluate and predict the actual maps conformity, Kappa coefficient and true skill statistic (TSS) were measured. The results showed that the highest values of kappa and TSS belong to the ANN (ê= 0.81, TSS= 0.8), MaxEnt (ê= 0.79. TSS= 0.57) and LR models (ê= 0.63, TSS= 0.55), respectively. Based on these results, it can be said that there is a strong relationship between model performance and the kinds of species distributions being modeled. Some methods performed generally better, but no method was superior in all circumstances.

Keywords: Artificial neural network, Logistic regression, Machine learning, Maximum entropy, True skill statistic

Abbreviations: **ANN:** Artificial neural networks; **GLM:** Generalized linear model; **LR:** Logistic regression; **MaxEnt:** Maximum entropy; **PVM:** Predictive vegetation modeling; **TSS:** True skill statistic

Introduction

Predictive vegetation modeling (PVM) can be defined as predicting the distribution of vegetation across a landscape based on the relationship between the spatial distribution of vegetation and certain environmental variables (Guisan and Zimmermann, 2000). It requires digital maps of the environmental variables, spatial information on the vegetation attribute of interest (e.g., species, type, abundance), usually from a sample of locations, and an appropriate statistical model (Phillips *et al*., 2006).

A variety of predictive vegetation modeling methods are available to predict potential suitable habitat for a species (Guisan and Thuiller, 2005; Zare Chahouki and Esfanjani, 2015). Generalized linear models (GLMs) have been used extensively in vegetation modeling research. When response data are binary, the appropriate GLM is a logistic model, which uses a logit link to describe the relationship between the response and the linear sum of the predictor variables (Guisan and Zimmermann, 2000; Hosmer and Lemeshow, 2000). MaxEnt is a maximum entropy based machine learning program that estimates the probability distribution for a species' occurrence based on environmental constraints. It requires only species presence data (not absence) and environmental variable (continuous or categorical) layers for the study area (Phillips *et al*., 2006).

One of the robust machine-learning approaches used in predictive vegetation modeling is artificial neural networks techniques (Thuiller *et al*., 2003). This has the advantage over other statistical techniques that it is more accurate and faster than other techniques when the problem is extremely complex, as well as, it does not require a prior knowledge of underlying process or assumptions of the structure of the target function, (Piri Sahragard and Zare Chahouki, 2015). Overall, different modelling approaches have the potential to yield substantially different predictions accuracy (Moisen and Frescino, 2002; Zare Chahouki *et al*., 2012). Despite all these, few studies have been conducted in order to compare different methods of modeling and specify each method capacity to anticipate in comparison with other methods in order to select the best modeling approach. The specific objectives of study were; i) develop models that describe the presence of four vegetation alliances by using logistic regression models, maximum entropy and artificial neural networks ii) compare model performance in terms of classification accuracy using Hosmer and Lemeshow, Area under the curve of receiver-operating characteristic (ROC) plots and mean of square error (MSE); iii) generate binary maps of each alliance and evaluation of predictive accuracy of the maps using Kappa and TSS.

Materials and Methods

Study area**:** The study area is located in the Khalajestan part of Qom province in Geographic coordinates area 50' 17º 0'' to 50' 24º 0'' E and 34' 40 º 30'' to 34' 43º 30'' N. This region is located in the west of Qom city of Iran and covers an area of 21,000 hectares (Fig. 1). Minimum and maximum altitudes in the study area are 1300 and 1700 meters above sea level, respectively.

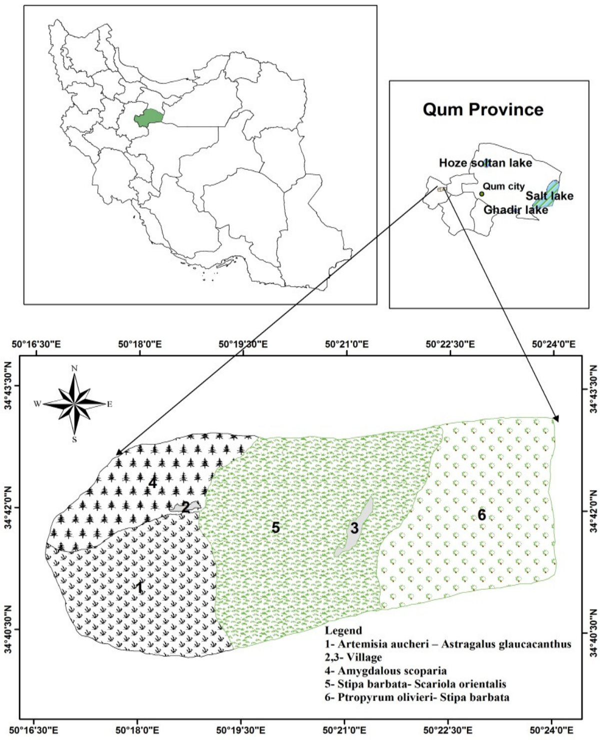

Fig 1. General location and vegetation types map of the study area.

Data collection: After determination of homogeneous units using basic maps of the study area (digital elevation, aspect, slope and geology maps, scale 1:25000), in the homogeneous units, vegetation sampling was carried out using random systematic method via the plots established along four transect with 150-200 m lengths. Depending on the plant species, the plot size was determined using minimal area method from 2 to 25 m^2. The sample size used in the study was determined to be 60 plots with respect to vegetation cover variations using statistical method. Besides vegetation data (name of plant species and canopy cover percent) information related to the geographical boundaries of habitats, slope, aspect and altitude were also recorded. For soil sampling, at each habitat, eight holes were drilled and samples were taken from 0-30 and 30-80 depths. After sampling, soil characteristics consisting gravel percent, texture, saturation moisture, available water, lime, gypsum, organic matter, acidity (pH), electrical conductivity (EC) and soluble solute (Na^+, Ca^{2+}, Mg^{2+}, K^+, Cl^-, Co^{32-}, HCO_3^- and SO_4^{2-}) were measured by standard methods. Using geostatistical and kriging interpolation method with the same spatial resolution (pixel size 30*30 meters) soil digital layers were prepared and stored in GIS. Arc GIS 9.3 and GS+ version 5 software were used for mapping soil properties. Digital elevation map of the region 1:25000 scale was used for mapping slope, aspect and elevation.

Model development**:** Modeling was performed using LR, MaxEnt and ANN approaches. In order to apply LR procedure, initial multicollinearity between variables was assessed and variables with variation inflation factors higher than 5 were eliminated (Fielding and Haworth, 1995). Then, relationships were extracted using SPSS version 18 and the model obtained was assessed using Hosmer and Lemshow Statistic (Hosmer and Lemshow, 2000). After providing maps of environmental variables in each of the models using geostatistic and Geographic information system (GIS), the variables coefficients entered in the model were multiplied by the corresponding layers in the GIS environment. Finally prediction maps of plant species habitat were obtained.

MaxEnt modeling was performed after preparation of the environment variable maps by using free software MaxEnt (http://www.cs.princeton.edu/~schapire/maxent/) which has been found to perform best among many different modeling methods. The area under the curve of receiver operating characteristic function (AUC) was used for evaluation of the discrimination ability (Fielding and Bell, 1997).The AUC ranges from 0.5 for an uninformative model to 1 for perfect discrimination. Also Jackknife analysis was used to determine the importance of variables. In order to perform ANN modeling, suitable combination of input variables of the input layer for each habitat was determined based on the results obtained from logistic regression. Moreover, to determine the optimal neural network structure in the middle layer, many networks with various topologies with adjustable parameters were implemented. Neural networks were built and trained with the neural network toolbox of MATLAB R2008 (The MathWorks Inc.). The best network was chosen using statistical criteria calculated in the test phase (*i.e.* MSE) and simulation was performed with the optimal network. After selecting the optimal networks of each habitat, this network was used to predict the probability of the presence or absence of each species in areas in which sampling had not taken place.

In order to determine optimal threshold, the current study applies sensitivity and specificity equal approach that is popular in ecology (Guisan *et al.*, 1998, Piri Sahragard *et al.*, 2015). After determining the optimal threshold for each plant species, the compliance between predictive and actual maps was evaluated through the calculation kappa. Also, due to dependence of the kappa measure on species prevalence, model evaluation was performed using true skill statistic (TSS). Like kappa, TSS takes into account both omission and commission errors ranging from -1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Allouche *et al.*, 2006).

Results and Discussion

***Models comparison*:** Evaluation of predictive models accuracy using HL, AUC and MSE statistics are given in table 1. Regarding Hosmer-Lemeshow results, the obtained equations are significant at one percent level. Considering AUC values and Sweet (1988) AUC classification, these results indicate a good predictive model accuracy of *Artemisia aucheri – Astragalus glaucacanthus* habitat and acceptable predictive model accuracy for *Pteropyrum olivieri- Stipa barbata, Scariola orientalis- Stipa barbata* and *Amygdalus scoparia* habitats. According to the MSE values, ANN models have high classification accuracy in the all habitats.

***Predictive accuracy evaluation of the maps using kappa and TSS*:** Evaluation of the correspondence between the actual and predictive maps reveals significant differences between predictive maps derived from the three methods (Table 2). This could be due to different quality models derived from the three methods. According to the LR results, conformity rate between prediction and actual map in different site varies from 0.42 (moderate level) to 0.91 (excellent level). Conformity rate of prediction maps with actual maps for *Amygdalus scoparia* (ê=0.91), *Pteropyrum olivieri- Stipa barbata* (ê=0.63) and *Stipa barbata -Scariola orientalis* (ê=0.58) habitats was excellent and good, respectively, but the accuracy of *Artemisia aucheri -Astragalus glaucacanthus* (ê=0.42) predictive map was down and the estimated conformity rate of prediction and actual maps was moderate. The outcome showed that LR model is capable to predict habitats distribution of *Amygdalus scoparia* with high accuracy, since these species has narrow amplitude (ê=0.91). In general, Logistic regression models provide a concise and probabilistic abstract of species environment relationships, as well as provide better specific model (Zare Chahouki *et al.*, 2010). Concerning the shape of the logistic regression function, *i.e.* a sigmoid curve, which is due to the nonlinear relationship between species and environmental factors, applying model is appropriate for these types of research (Zare Chahouki and Khalasi Ahwazi, 2012).

According to the results of maximum entropy approach, level of agreement of predictive maps at each site, show an excellent correspondence for actual and predictive maps of *Artemisia aucheri–Astragalus glaucacanthus* (ê=0.91), moreover, it shows that predictive maps of *Amygdalus scoparia*, *Pteropyrum olivieri- Stipa barbata* and *Stipa barbata -Scariola orientalis* have very good correspondence with the actual maps (Table 2). Based on the results obtained, the maximum entropy method ranked after ANN method in terms of performance. In agreement with earlier studies on species distributions using this approach, MaxEnt well performed especially

Table 1. Statistics of the models discrimination ability to predict the presence of plant species

Vegetation type	HL	AUC	MSE
Pteropyrum olivieri- Stipa barbata	1	0.73	0.00017
Scariola orientalis- Stipa barbata	0.99	0.75	0.0078
Artemisia aucheri- Astragalus glaucacanthus	0.99	0.93	0.00054
Amygdalus scoparia	0.99	0.89	0.00027

when modelling distributions of species with a greater area of occupancy and MaxEnt is the preferred alternatives when fundamental niche (different from occupied niche) is more important (Philips *et al*., 2006; Hosseini *et al*., 2013). As well as maximum entropy method is a generative method in contrast to GLM which is regarded as diagnostic methods, and can provide better predictions when the training data are limited (Ng and Jordan, 2001).

Artificial neural network results showed excellent correspondence for predictive habitat map of *Pteropyrum olivieri- Stipa barbata*, also predictive maps of *Amygdalus scoparia*, *Artemisia aucheri – Astragalus glaucacanthus* and *Stipa barbata-Scariola orientalis* have very good correspondence with actual maps (Fig. 2). Several studies show that artificial neural network approach has better performance than other methods. in addition, this method has specifically superior performance compared to regression methods (Zare Chahouki and Khalasi Ahvazi, 2012). Overall, it can be said that ANN method could be a more valid alternative than spatial statistical methods due to the ability of a neural network approach in modeling nonlinear relationships between variables and phenomena. However, it should also be considered that this method has also some errors and uncertainties which must be considered by its users (Piccinini, 2011).

Table 2. TSS and kappa statistic values obtained for each vegetation type by the used methods

Vegetation type		Models		
		LR	MaxEnt	ANN
Pteropyrum olivieri-	Kappa	0.63	0.68	0.90
Stipa barbata	TSS	0.56	0.53	0.84
Stipa barbata -	Kappa	0.58	0.79	0.72
Scariola orientalis	TSS	0.34	0.59	0.63
Artemisia aucheri–	Kappa	0.42	0.91	0.78
Astragalus glaucacanthus	TSS	0.16	0.58	0.84
Amygdalus scoparia	Kappa	0.91	0.80	0.85
	TSS	0.51	0.61	0.90

***Model assessment and comparison of methods based on test data*:** As noted earlier, for a more accurate assessment of the used models and considering the fact that kappa coefficient is a criterion related to the prevalence species (Alloche *et al*., 2006), TSS was applied to evaluate the models obtained. Model comparisons based on sensitivity and specificity of the models, indicate that the highest values of Kappa and TSS belong to the ANN, MaxEnt and LR models, respectively.

Based on the sensitivity values obtained for each of the models, the logistic regression (LR) models for the *Artemisia aucheri -Astragalus glaucacanthus* alliance was the poorest model because none of the probability of presence values exceeded 0.33. However, the proposed model for *Pteropyrum olivieri- Stipa barbata* is the most accurate model to classify the presence or absence of species in this habitat (Sensitivity 0.79). The poorest maximum entropy model is related to *Artemisia aucheri-Astragalus glaucacanthus,* since considering the 0.2 optimal thresholds model ability has not exceeded 0.45 in diagnosis of the presence or absence of plant species. The poorest and most powerful models of artificial neural network with 0.68 and 0.96 diagnostic ability are related to *Stipa barbata -Scariola orientalis* and *Amygdalus scoparia* habitats, respectively (Table 3).

Predictive species distribution modeling can provide a valuable tool for conservation planning and biodiversity management, especially in poorly surveyed regions that are under accelerating pressure of habitat loss and degradation (Austin, 2002; Araújo and Guisan, 2006). At first sight in the current research, the overall performance of all models seems to be rather equal. But overall assessment of the used methods by Kappa and TSS statistics showed that there is significant difference in model performance in terms of both measures considered (Kappa index and TSS). According to kappa and TSS values, artificial neural networks performed generally better (k= 0.81, TSS= 0.8), immediately followed by MaxEnt (k= 0.79, TSS= 0.57) especially when modelling distributions of species with a greater area of occupancy and LR was the lowest (k= 0.63, TSS= 0.55).

Conclusion

In general, it can be said that LR is a good alternative in case the ecological niche of the species is narrow. While MaxEnt is substantially superior to predict geographical distributions of plant species when fundamental niche (different from occupied niche) is more important and can perform well with fairly few examples due to employing regularization. In general, our results encourage further use of MaxEnt for species distribution modeling, especially when the sample size is small. In addition, ANN method could be a more valid alternative than spatial statistical methods due to the ability of a neural network approach in modeling nonlinear relationships between variables and phenomena.

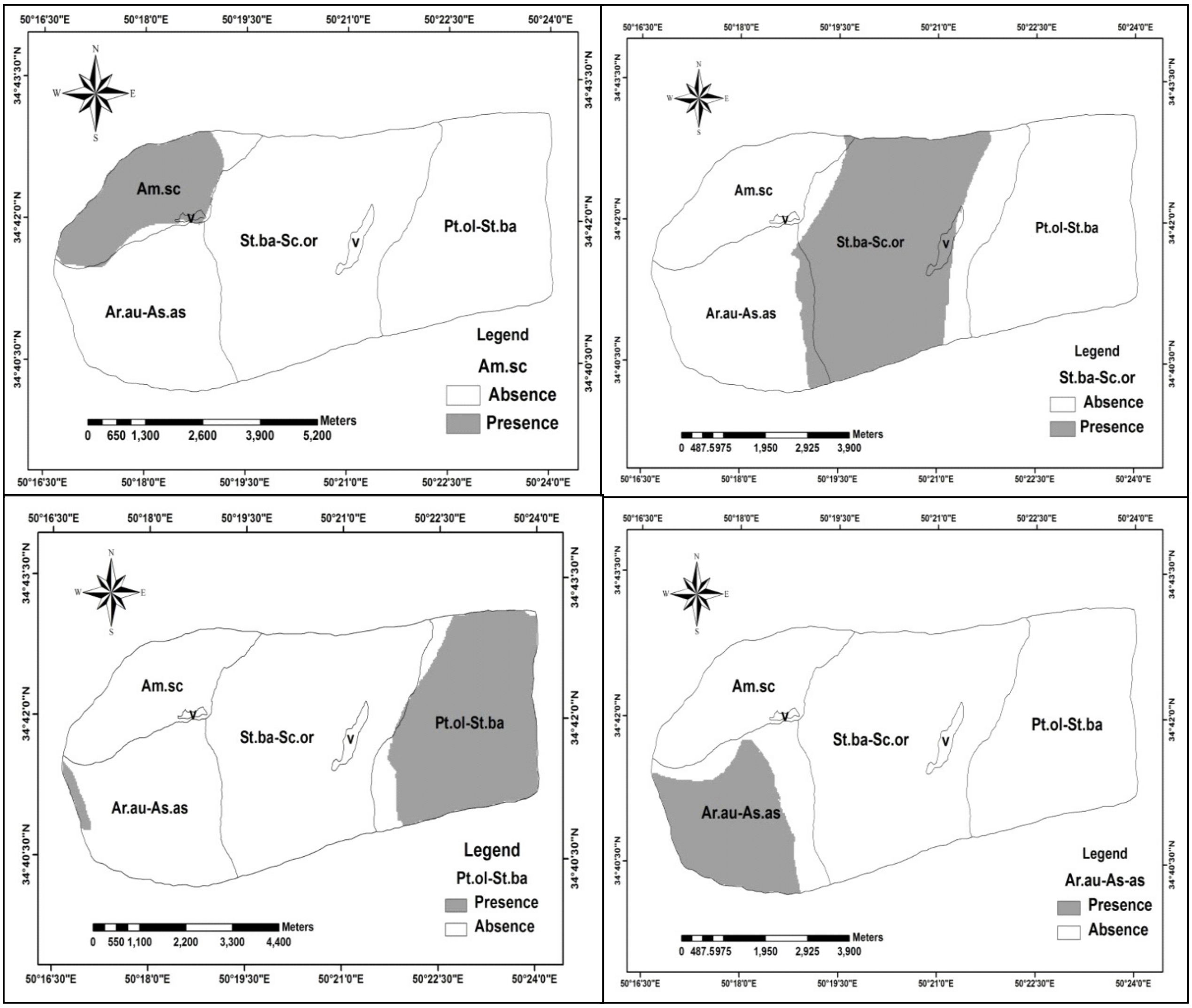

Fig 2. Actual and predicted species distribution maps (Predictive maps shown darker).

Table 3. Optimum probability data threshold and sensitivity/specificity for all models based on test data

Vegetation type	**Model**					
	LR			**MaxEnt**		
	Optimum probability	**Sensitivity**	**Specificity**	**Optimum probability**	**Sensitivity**	**Specificity**
Pteropyrum olivieri- Stipa barbata	0.5	0.79	0.77	0.2	0.56	0.97
Stipa barbata -Scariola orientalis	0.3	0.50	0.84	0.5	0.68	0.81
Artemisia aucheri – Astragalus glaucacanthus	0.3	0.33	0.83	0.2	0.45	0.99
Amygdalus scoparia	0.3	0.68	0.83	0.3	0.64	0.97

Vegetation type	**Model**		
	ANN		
	Optimum probability	**Sensitivity**	**Specificity**
Pteropyrum olivieri- Stipa barbata	0.2	0.85	0.99
Stipa barbata -Scariola orientalis	0.3	0.68	0.95
Artemisia aucheri – Astragalus glaucacanthus	0.5	0.91	0.93
Amygdalus scoparia	0.7	0.95	0.95

References

Allouche, O., A. T. Soar and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa, and the true skill statistic (TSS). *Journal of Applied Ecology* 43: 1223-1232.

Araújo, M.B. and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33: 1677-88.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modeling* 157: 101-118.

Fielding, A.H. and P. F. Haworth. 1995. Testing the generality of bird-habitat models. *Conservation Biology* 9: 1446-1481.

Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38-49.

Guisan, A., J. P. Theurillat and F. Kienast. 1998. Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science* 9: 65-74.

Guisan, A. and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modeling* 135: 147-186.

Guisan, A. and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8: 993-1009.

Hosmer, D.W and S. Lemeshow. 2000. *Applied Logistic Regression*. Wiley, New York. pp. 307.

Hosseini, S. Z., M. Kappas., M. A. Zare Chahouki., G. Gerold., S. Erasmi. and A. Rafiei Emam. 2013. Modelling potential habitats for Artemisia sieberi and Artemisia aucheri in Poshtkouh area, central Iran using the maximum entropy model and geo-statistics, *Ecological Informatics* 18: 61-68.

Moisen, G. G. and T. S. Frescino. 2002. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modeling* 157: 209-225.

Ng, A.Y. and M. I. Jordan. 2001. On discriminative versus generative classifiers: a comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems* 14: 605-610.

Phillips, S. J., R. P. Anderson and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modeling* 190: 231-259.

Piccinini, C. 2011. *Assessing the impact of climate change on plant distributions using Artifical Neural Networks*. Kingston University, London. pp. 129.

Piri Sahragard, H. and M.A. Zare Chahouki. 2015. An evaluation of predictive habitat models performance of plant species in Hoze soltan rangelands of Qom province. *Ecological Modelling* 309-310: 64-71.

Piri Sahragard, H., M. A. Zare Chahouki and H. Gholami. 2015. Predictive distribution models for determination of optimal threshold of plant species in central Iran. *Range Management and Agroforestry* 36: 146-150.

Sweet, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240: 1285-1293.

Thuiller, W., M.B. Araújo. and S. Lavorel. 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14: 669-80.

Zare Chahouki, M. A., H. Azarnivand., M. Jafari. and A. Tavili. 2010. Multivariate statistical methods as a tool for model based prediction of vegetation types. *Russian Journal of Ecology* 41: 84-94.

Zare Chahouki, M.A., L. Khalasi Ahvazi and H. Azarnivand. 2012. Comparison of three modeling approaches for predicting plant species distribution in mountainous scrub vegetation (Semnan rangelands, Iran). *Polish Journal of Ecology* 60: 105-117.

Zare Chahouki, M.A. and L Khalasi Ahvazi. 2012. Predicting potential distributions of Zygophyllum eurypterum by three modeling techniques ENFA, ANN and logistic in North East of Semnan Iran. *Range Management and Agroforestry* 33: 68-82.

Zare Chahouki, M. A. and J. Esfanjani. 2015. Predicting potential distribution of plant species by modeling techniques in southern rangelands of Golestan, Iran. *Range Management and Agroforestry* 36: 66-71.